公開

研 究 終 了 報 告 書

「メタゲノムビッグデータを活用した微生物の環境適応因子の解明」

研究期間:　2021 年 10 月～ 2024 年 3 月
研 究 者: 西村 陽介

1. 研究のねらい

微生物は極限環境を含む地球上の様々な環境へ適応し、繁栄してきた。近年、微生物の環境適応の鍵は遺伝子レパートリーの刷新にあり、それによって微生物の進化が駆動されてきたと提唱されている。特に、遺伝子水平伝播は微生物の環境適応において重要であり、「環境適応の鍵を握る遺伝子」(つまり、その環境での生存に必要な因子、または競争優位性を築くための因子)は、その環境に生息する様々な生物系統に、遺伝子水平伝播を介して広まっていると推測される。しかし、微生物の様々な環境への適応・進化の過程についての理解は乏しく、「環境適応の鍵を握る遺伝子」はどのようなものか、といった微生物の環境適応の全体像は謎に包まれている。また、共生菌・病原菌・ウイルスでは、環境とともに宿主への適応度を上げることが「環境適応」であるが、その過程への理解も乏しい。環境適応の全体像の把握には、大規模メタゲノムデータを用いて、「遺伝子」と「環境」の関係性を網羅的に解析し、知識化することが有効である。しかし、そのための方法論は未発達である。

次世代シーケンサーの普及とメタゲノムデータの増加により、機能未知遺伝子も増加の一途をたどっており、微生物遺伝子のうち 40-60%は機能未知遺伝子であるとの推定がある。遺伝子の応用利用には、まず遺伝子機能を推定することが重要であるが、機能未知遺伝子は配列類似性などの従来の手法では機能推定ができない。そこで、有用な遺伝子資源を、環境メタゲノムから効率よく発見するためには、遺伝子の環境での分布の情報を有効活用するための手法開発が必要である。

本研究は、「各メタゲノムが属する環境を分類する手法」を含む、様々な情報学的手法の開発を通じて、メタゲノムデータを環境横断的・網羅的に解析するための手法を確立する。そして、公開されているメタゲノム・ビッグデータを利用してデータ駆動型研究を実施し、各遺伝子の環境特性を知識化し、微生物の環境適応因子の全体像を解明することで、生命進化への新しい視点を切り開く。また、遺伝子の環境特性や遺伝子水平伝播の網羅的な情報を活用することで、機能未知遺伝子の機能を推定し、有用な遺伝子資源を探索する。さらに、知識化された遺伝子情報基盤を公開し、「遺伝子と環境の関連性についての情報リソース」として世界的な普及を推進することで、微生物関連科学の発展に貢献する。

2. 研究成果

(1)概要

「世界最大級のメタゲノム情報基盤の構築」と、「系統樹ベースの遺伝子解析手法の開発」に成功した。それを土台として多数の成果が創出されており、数多くの共同研究が展開されている。

「世界最大級のメタゲノム情報基盤の構築」については、様々な環境に由来する公共メタゲノムデータの取得と解析を行った。18390 メタゲノムサンプル(253 兆塩基)でアセンブリ等の解

析を完了しており、現段階で世界最大級のメタゲノム情報基盤となっている。また、本研究の全てのメタゲノムが出版済み論文（n=956）に紐づき、そのメタデータを収集している点で、競合するメタゲノム情報基盤に対する優位性がある。本情報基盤を土台として、海洋微生物のゲノム多様性と、その環境適応の鍵となる因子の解明を推進した。海洋メタゲノムから 5 万以上の原核生物ゲノムを解読し、世界最大の海洋生物ゲノムカタログを構築することで、様々な海洋環境でのゲノム多様性を明らかにし、多数の国際・国内共同研究に発展している。特に、新発見の窒素固定菌を含む、北極海に生息する窒素固定菌のゲノム解析によって、極域の環境への適応の鍵となる遺伝子を推定した。また、海洋光合成細菌のゲノムと環境適応の関係性を解析し、環境下で入手可能な栄養源が分布に大きな影響を及ぼし、ゲノム進化に影響を与えていることを明らかにした。また、全く新しい系統 Woeseiales の光合成細菌を複数同定した。

「系統樹ベースの遺伝子解析手法の開発」については、2つの解析パイプラインの開発に成功し、遺伝子オーソログの大規模系統樹の構築と、その環境中での分布の定量化が自動化された。具体的には、（1）大規模タンパク質 DB から、各オーソログ遺伝子に相当する配列を網羅的に収集し、リファレンス系統樹の構築、各クレードの同定までを自動化した。（2）メタゲノム中タンパク質配列を任意の遺伝子のリファレンス系統樹にマッピングし、各クレードの存在頻度の定量を自動化した。本手法は、遺伝子の進化と環境適応を理解し、有用な遺伝子資源を応用利用する共同研究の展開に広がっている。例えば、マグネトソーム遺伝子クラスターの探索、光受容体 pyp の探索とその応用利用、環境に由来するプラスミド複製開始因子 rep の網羅的探索と情報基盤の確立などの共同研究を展開している。

（2）詳細

【1．メタゲノムデータ取得と環境情報の整備】

メタゲノム配列の取得については、NCBI データベースから入手可能な、地球上の様々な環境に由来するメタゲノムデータ 18390 サンプル（253 兆塩基）分を取得し、リード配列の品質フィルタリング、配列アセンブリ、遺伝子予測などの解析を行った。順次新しくメタゲノムデータが公開されるため、新しいデータも取得して同様の解析を行ってきた。これらのメタゲノムデータセットは、全てのサンプルが査読付き論文で公開されており、論文からも環境データを収集している点が、海外の競合データベース（欧 EBI MGnify や米 JGI IMG/M）とは異なる、本研究独自の優位性である。また、概要に記したように、海洋環境のメタゲノムについては、2057 メタゲノムから 52325 個の原核生物ゲノムを構築し、「OceanDNA MAG カタログ」として公開し、様々な海洋環境でのゲノム多様性を明らかにした（論文1）。本ゲノムカタログから多数の研究成果が生まれている（論文2など）。

環境メタデータの収集については、データベース上で分類されている、海水・淡水・土壌などの大きなカテゴリに加えて、温度、水深、塩分濃度等の環境メタデータや、サンプリング後のフィルタリングの有無やそのフィルターサイズ、またインキュベーションの有無などの情報について、データベースや論文から収集し、エクセルファイルの形で情報を整備した。作業の手順としては、臨時研究補助員3名が情報を収集してファイルにまとめ、研究代表者が各項目の確認を行っている。現段階で 5000 以上のサンプルで、研究代表者による確認が終わった。淡水や海水については、水温や塩分濃度といった項目の情報は比較的得やすい傾向にある。その一方で、土壌や堆積物等については液体サンプルよりも環境データの計測が困難であり、例え

ば直上の水（液体）に関するメタデータが記載されているなど、直接的に環境情報を文献等から得ることは難しい場合が多いことが分かった。その場合でも、論文中には得られた環境に関する詳細が記されており、各サンプルが論文に紐づいていることで、必要に応じて論文を参照できることは、本情報基盤が持つ他のデータベースにはない優位性である。

環境クラスターの構築については、データベースから得られる海洋・淡水などのカテゴリは、分類単位としては大きすぎて、それぞれのカテゴリ内に大きな環境多様性が存在する。また、テキスト化された詳細な環境情報は、「情報の粒度が揃わない」・「必ずしも情報が手に入らない」・「全ての論文から情報を取得する手間をかけられない」という問題がある。これらの点を克服し、高精度な環境分類を行うため、メタゲノム配列を活用して環境情報を整備する。具体的には、各メタゲノムに含まれる配列の特徴をもとにして、メタゲノム間の距離を定義し、この距離に基づいて、メタゲノムを階層クラスタリングすることで、階層性を持つ環境クラスターを定義する。利用する特徴の候補としては、(1)各遺伝子オーソログの頻度、(2)系統組成、(3)k-mer組成が考えられる。現在までに、各メタゲノムにおけるこれらの特徴量を算出しており、(3)に関してはメタゲノム間の距離計算を終えている。今後、これらの特徴量をベースに階層クラスタリングを行い、環境メタデータと照合することでこの推定手法の妥当性を評価し、最良の方法を選択する。また、得られた環境クラスターには、環境メタデータを紐付けることで、それぞれの環境クラスターを特徴づける。

【2. オーソログ情報とリファレンス系統樹の整備】

オーソログ情報の整備については、既存のタンパク質オーソログ（または機能ドメイン）のデータベース（KEGG KO、Pfam、TIGRFAMs など）から提供されている HMM（各アラインメントに関する隠れマルコフモデル）を非冗長化した。ただし以下の解析においては、系統樹計算にかかる時間を短縮するため、情報が最も充実しており代謝解析等で最も良く用いられている原核生物に関する KEGG KO を中心に解析を進めた。

リファレンス系統樹の整備については、公開されている微生物ゲノムに含まれる各遺伝子オーソログを UniParc データベース（2021 年 12 月版、4.6 億タンパク質配列）を用いて網羅的に収集した。この配列全体に対して KEGG KO に対する相同性検索を行い、それぞれのオーソログ遺伝子セットに相同性を示す配列をまとめ、80%の配列相同性で代表配列を取り出すことで冗長性を除いた。この代表配列をもとに配列アラインメントを取り、リファレンス系統樹を計算中である。配列アラインメントに想定以上の時間を要したため、非常に多数の配列を含む場合は、相同性の高いグループごとにアラインメントを取得してから、それらを一つのアラインメントにまとめ上げる手法によって、大幅に計算時間を短縮することに成功した。また、系統樹の計算には FastTree を使用した。また、各系統樹の樹形とその信頼性を判断材料として、クレードの分割を自動化する手法を開発した。今後、それぞれの遺伝子（クレード）の生物系統の分布についても知識化する。

【3. リファレンス系統樹の環境特性を知識化】

各メタゲノムでの定量化については、各メタゲノムから得られた遺伝子群を、リファレンス系統樹にマッピングすることによって、各遺伝子（クレード）の存在頻度を定量化する。この手法については、解析パイプライン「PPP」を開発し、メタゲノム由来のタンパク質配列を入力とする

ことで、当該遺伝子に相同性のある配列を検出し、クレードごとに定量化する解析を自動で行うことが可能となった。各遺伝子オーソログ単位での定量化については約 1 万サンプルで解析が完了した。各クレード単位でのより詳細な定量化は、【2】の完了後に進める。

環境クラスターでの定量化については、【1】で定義した環境クラスターを活用し、それぞれの遺伝子（クレード）が、各環境クラスターでどの程度の頻度で見られるかを定量化する。その結果に基づいて、遺伝子（クレード）の環境特異性の高さを指標化する手法を開発することで、遺伝子（クレード）の環境特性を知識化する。

【4. 環境適応の「鍵」遺伝子の推定】

本項目は、上記【3】の解析終了後に行う予定である。本項目ではまず、遺伝子水平伝播の推定を行う。各遺伝子（クレード）において、各配列の生物系統情報から、水平伝播の痕跡を推定する。特に、特定の環境に偏って存在する遺伝子（クレード）において、生物ドメインや門といった、大きな分類をまたぐ水平伝播が起きているかを重点的に探索し、情報の整備を行う。そのために、【1】で定義した環境クラスター、及び【2】で整備した階層的なクレード情報を活用する。また、生物系統情報の取得に関しては、研究代表者が開発した様々なメタゲノム解析パイプラインを活用し、多くの配列に関して生物系統を推定する。

また、共起遺伝子情報の整備を行う。各遺伝子（クレード）に関して、ゲノム上で近傍に存在する組み合わせを抽出し、共起遺伝子として知識化する。メタゲノムに由来する遺伝子については、遺伝子が含まれるコンティグの情報を活用する。UniParc 等の配列データベースに由来する遺伝子についても、ゲノム情報やコンティグ情報を辿ることによって、近傍遺伝子の情報を整備する。

環境適応の「鍵」となる遺伝子の推定には、上記で整備された遺伝子（クレード）の環境特異性・生物系統の多様性・遺伝子水平伝播・共起遺伝子などの情報を統合的に用いて推定を行う。推定方法の妥当性については、よく調べられている遺伝子の既知の知見と照合することで評価する。遺伝子（クレード）の環境特異性については、3の統計学的検定の結果を用いて、各遺伝子（クレード）について最も有意に多く存在する環境クラスターの情報から判断する。また、生物系統の多様性・遺伝子水平伝播・共起遺伝子については、上記で整備した情報を活用する。

【5. 有用遺伝子の探索】

概要に記したように、メタゲノムデータからの網羅的な遺伝子配列探索により、個々の遺伝子の進化と環境適応を理解し、有用な遺伝子資源の応用利用を促進するための共同研究を多数実施している。今後、機能未知遺伝子の環境特性・共起遺伝子・生物系統分布を総合的に解析することで、機能未知遺伝子から有用遺伝子を探索するための「一般化」された方法論の構築を目指す。特に、環境特異性の高い機能未知遺伝子に注目し、探索対象となる遺伝子に応じて総合的に吟味する。探索の方法は、対象遺伝子に依存して柔軟に対応できる形で設計する。例えば、新規光受容体の探索には、光の強く当たる環境に偏在し、近傍に光受容体関連遺伝子があるものを選抜する。また、共同研究を通じて、選抜された候補遺伝子に関して、大腸菌を用いた異種発現系などにより機能解析実験を行い、機能や活性を確認する。

公開

3. 今後の展開

本研究では世界最大規模のメタゲノム情報基盤を構築し、系統樹ベースの遺伝子解析手法を開発した。この成果を活用した共同研究により、個々の遺伝子ファミリーの配列多様性を探索し、その由来の環境の情報を酵素設計等の応用利用に活用する取り組みを行っている。今後も、本研究の成果を最大限に活用して、微生物の環境適応プロセスを理解し、遺伝子の応用利用のための研究を進めていく。

本研究は生命情報学（バイオインフォマティクス）を軸にした研究開発であり、成果物はデータベースや解析ツールの形で一般公開する。2-3 年以内に全ての成果を発表することで、他の研究者が自由に使用可能にする。公開したメタゲノム情報基盤とその関連ツールによって本研究の成果が活用され、幅広い分野の研究者が大規模メタゲノムデータを直感的に利用可能にし、「環境とバイオテクノロジー」の関連分野おいて科学的な発見や応用利用を加速させることで、本研究の社会実装を達成する。

生命情報学の分野では、研究の国際化が加速しており、例えばゲノムに基づく微生物分類体系や、大規模ゲノムデータリソースの構築については、様々な国の研究者が議論し、共同で構築していくことが一般的となりつつある。今後、本研究で得られた成果を継続的に発表していくことで、研究代表者が日本のメタゲノム研究を盛り上げて国際的な舞台に立ち、共同研究や研究コンソーシアムに参画することで、大規模メタゲノムデータを活用した微生物研究を先導していきたい。

4. 自己評価

本研究では、世界最大級のメタゲノム情報基盤を構築し、系統樹ベースの遺伝子解析手法の開発に成功した。特に海洋環境に関しては、メタゲノム情報基盤を活用して、微生物ゲノムや遺伝子の獲得とその環境適応との関係性についての新しい知見が次々と得られている。それによって、本研究の目的である微生物の環境適応メカニズムの解明が飛躍的に進展した。一般社会における地球環境保全への危機感が高まっている中で、この成果に関する論文発表と2件の新聞報道によって、「微生物と環境の密接な関係」を解明することの重要性について、紙面で訴える機会を得た。

研究計画の進捗としては、環境メタデータ収集や網羅的な系統樹の計算に関して予定したスケジュールでは進まなかった面もあるが、本研究の研究費により臨時研究補助員３名を雇用することで、環境メタデータ収集はかなり加速された。作業労力を分担することが比較的難しいバイオインフォマティクス分野において、環境メタデータ収集作業への助力を研究費によって得たことは、本研究にとって大変有意義であった。

本研究で開発したメタゲノム情報基盤や解析パイプラインを活用して、多数の成果を上げたことにより、国際的な競争が活発である「大規模メタゲノム解析」の分野で国内を代表する研究を展開することができたと自負している。また、海洋5万ゲノムの成果発表をきっかけとして、国内外を問わず多数の共同研究が始動しており、本研究が国内外の研究者に与えたインパクトは大きいと考えられる。本研究によって、研究代表者が国際的に認知され、今後も活躍していくための研究基盤が整備されたと考えている。

公開

5. 主な研究成果リスト

（1）代表的な論文（原著論文）発表

研究期間累積件数： 3件

1. *Yosuke Nishimura, and Susumu Yoshizawa. ‘The OceanDNA MAG Catalog Contains over 50,000 Prokaryotic Genomes Originated from Various Marine Environments.’ *Scientific Data* (2022) 9:305. doi:10.1038/s41597-022-01392-5 （*は責任著者）

海洋には「微生物ダークマター」と呼ばれる未知の微生物が数多く存在する。本研究はメタゲノムデータから個々の微生物ゲノムを解読する手法（MAGRE）を開発し、公共データベースにある、様々な海洋環境に由来する大規模メタゲノムデータ（約 29 兆塩基対）を解析した。59 門にまたがる 5 万以上の原核生物ゲノムを解読し、世界最大の海洋微生物ゲノムカタログ「OceanDNA MAG カタログ」として公開した。本カタログは、海洋微生物の生態や進化の過程の解明に貢献し、地球環境保全のための重要な知識基盤となる。

2. #Takuhei Shiozaki, #Yosuke Nishimura, Susumu Yoshizawa, Hideto Takami, Koji Hamasaki, Amane Fujiwara, Shigeto Nishino, Naomi Harada. ‘Distribution and survival strategies of endemic and cosmopolitan diazotrophs in the Arctic Ocean.’ *The ISME Journal* (2023) 17:1340–50 doi:10.1038/s41396-023-01424-x （#は共に筆頭著者）

上記の OceanDNA MAG カタログから、北極海に分布する計 7 種の未培養窒素固定生物ゲノムを同定し、その全球分布パターンから北極固有種と、普遍種に分類した。ゲノム解析により、北極固有種のみが持つ特殊な遺伝子群が同定され、この遺伝子群により北極の特殊な環境に適応していることが示唆された。一方、普遍種は低温適応において機能することが知られている遺伝子を持っていた。普遍種はこの低温適応性によって全球(深海や極域の冷水でも)に存在していると考えられる。本成果は海洋窒素固定生物の地球上での分布と生態に関する新たな視座を与えた。

3. *Yosuke Nishimura, Kohei Yamada, Yusuke Okazaki, *Hiroyuki Ogata. 2024. ‘DiGAlign: versatile and interactive visualization of sequence alignment for comparative genomics’ Microbes and Environments 39(1):ME23061. doi:10.1264/jsme2.ME23061 (*は責任著者)

シンテニーマップ表示機能により、迅速な比較ゲノム解析ができるウェブツール DiGAlign を開発した。DiGAlign は、一度に最大 300 ゲノムの大規模な比較ゲノム解析が可能であり、ゲノム系統樹である「ガイドツリー」や、遺伝子の同定とその機能予測など、独自の機能を多数備えている。これらの機能により、仮説や事前知識なしにゲノムデータを解釈することで、大量のゲノム情報を入手可能な現代におけるゲノム解析のスピードを加速する。DiGAlign は微生物の大部分を占め、「微生物ダークマター」などと呼ばれる未培養微生物が持つ物質生産・エネルギー変換の潜在能力や、その進化、生態学的な役割等の解明を推進する。

（2）特許出願

研究期間全出願件数： 0 件（特許公開前のものも含む）

公開

（3）その他の成果（主要な学会発表、受賞、著作物、プレスリリース等）

【学会発表】

1. Yosuke Nishimura, Susumu Yoshizawa

“The OceanDNA MAG catalog: an unprecedented-scale genome resource of marine prokaryotes”

ASME12 (Asian Symposium on Microbial Ecology)

2022 年 4 月 18 日 国際学会・招待講演

【プレスリリース】

2. 「微生物ダークマター」を解き明かす —世界最大の海洋微生物ゲノムカタログ—

2022 年 6 月 17 日　東京大学 大気海洋研究所、科学技術振興機構（JST）

（代表的な論文（原著論文）発表1.のプレスリリース）

3. 北極海に生息する窒素固定生物のゲノム解読に成功 —北極固有種の存在とその特徴が明らかに—

2023 年 5 月 24 日　東京大学、海洋研究開発機構、科学技術振興機構（JST）

（代表的な論文（原著論文）発表2.のプレスリリース）

【新聞報道】

4. 「微生物ダークマターに迫る」（連載 ぶらっとラボ）

2022 年 7 月 25 日、朝日新聞夕刊3面

（代表的な論文（原著論文）発表1.の紹介記事）

5. 「北極海の窒素固定生物 ゲノム解読成功」

2023 年 6 月 2 日、科学新聞4面

（代表的な論文（原著論文）発表2.の紹介記事）